BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68 — A2

(11) International Publication Number: WO 95/17522

(43) International Publication Date: 29 June 1995 (29.06.95)

(21) International Application Number: PCT/GB94/02789

(22) International Filing Date: 21 December 1994 (21.12.94)

(30) Priority Data:
9326052.9 21 December 1993 (21.12.93) GB

(71) Applicant: UNIVERSITY OF LEICESTER [GB/GB]; University Road, Leicester LE1 7RH (GB).

(72) Inventors: JEFFREYS, Alec, John; 14 Asquith Boulevard, Leicester LE2 6FA (GB). ARMOUR, John; 35 Goods Lane, Syston, Leicestershire LE7 2JL (GB).

(74) Agents: McNEIGHT, David, Leslie et al.; McNeight & Lawrence, Regent House, Heaton Lane, Stockport, Cheshire SK4 1BS (GB).

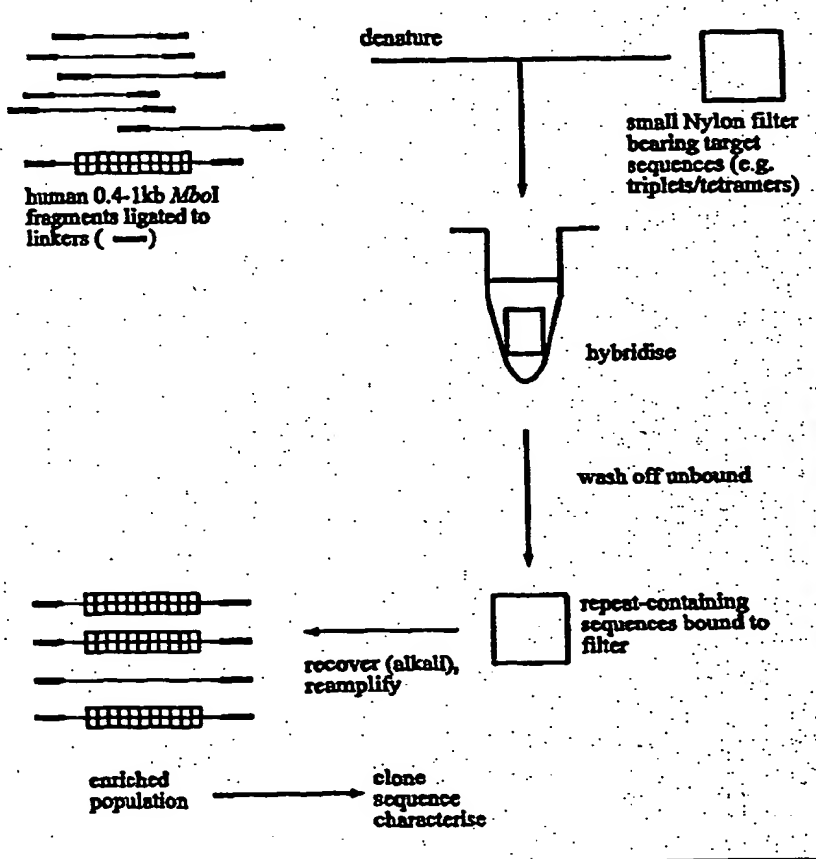
(81) Designated States: AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, HU, JP, KP, KR, KZ, LK, LU, MG, MN, MW, NL, NO, NZ, PL, PT, RO, RU, SD, SE, SK, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: IDENTIFICATION OF SIMPLE TANDEM REPEATS

(57) Abstract

The present invention concerns a method for the identification from DNA of a fragment comprising a simple tandem repeat locus comprising the steps of: i) contacting a DNA library with at least one hybridisation probe so as to identify a population of DNA fragments enriched for simple tandem repeats; ii) isolating and cloning said population; and iii) screening of the resulting DNA library so as to identify an individual fragment comprising a simple tandem repeat locus. Also provided are simple tandem repeats isolated by the method of the present invention, characterised in that they may be amplified at least in part by PCR using a specified pair of primers, together with amplification primers and probes specific to the simple tandem repeats so isolated. The present invention also provides methods of genetic characterisation using the aforementioned simple tandem repeats, primers and probes.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria
AU	Australia
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CS	Czechoslovakia
CZ	Czech Republic
DE	Germany
DK	Denmark
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgystan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakhstan
LI	Liechtenstein
LK	Sri Lanka
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SI	Slovenia
SK	Slovakia
SN	Senegal
TD	Chad
TG	Togo
TJ	Tajikistan
TT	Trinidad and Tobago
UA	Ukraine
US	United States of America
UZ	Uzbekistan
VN	Viet Nam

IDENTIFICATION OF SIMPLE TANDEM REPEATS

The present invention concerns methods for the identification from DNA, in particular from genomic DNA, of a fragment comprising a Simple Tandem Repeat (STR) locus, together with simple tandem repeat loci, primer sequences and hybridisation probes, as well as methods of genetic characterisation using the aforementioned simple tandem repeats, primers and probes.

Hybridisation techniques have been used in the past as preparative steps in the selection from cDNA libraries of sequences hybridising with cloned genomic DNA (see Parimoo, S., *et al.* (1991). *Proc. Nat. Acad. Sci. U.S.A.* 88: 9623-9627; and Lovett, M., *et al.* (1991). *Proc. Nat. Acad. Sci. U.S.A.* 88: 9628-9632) and in the isolation of (AC)n dinucleotide repeats from the mouse genome using immobilised short oligonucleotides as the hybridisation "target" (see Karagyasov, L., *et al.* (1993). *Nucleic Acids Res.* 21: 3911-3912). However, these techniques have various disadvantages, a primary disadvantage being that the use of short oligonucleotide "targets" results in hybridisation with a relatively restricted range sequences due to the inability of the "target" oligonucleotides to tolerate mismatches.

This inability to tolerate mismatches in the screening of libraries using relatively short oligonucleotides composed of perfect repeats (see, for example, Li, S.-H., *et al.* (1993). *Genomics* 16: 572-579) is further exemplified by the fact that loci containing frequently interspersed repeat unit variants may not be reliably detected. Interspersion of different repeat unit types is a common feature of many highly variable minisatellite loci, and has been exploited in the analysis of the mechanisms involved in the evolution of these longer repeat loci (see, for example, Armour, J.A.L., *et al.* (1993). *Human Mol. Genet.* 2: 1137- 1145).

Hybridisation screening of unenriched genomic libraries in the past has been successful in isolating simple tandem repeat loci from the human genome (see, for example, Weissenbach, J., *et al.* (1992). *Nature* 359: 79) but suffers from the general disadvantage of inefficiency; very large numbers of clones need to be screened from small-insert libraries for each positively hybridising clone, and large-insert (cosmid) clones require subsequent subcloning or other manipulation (see, for example, Edwards, A., *et al.* (1991). *Am. J. Human Genet.* 49: 746-756; and Lagerstrom, M., *et al.* (1991). *PCR Meth Appl* 1: 111-119) to determine sequence immediately adjacent to the repeat block.

The present invention overcomes the aforesaid problems of the prior art and provides a simple and efficient method for the isolation of simple tandem repeat loci from DNA libraries and in particular from genomic DNA. This method is based upon prior enrichment for tandemly repeated DNA fragments, a prior enrichment which is sensitive to the presence of tandemly repeated DNA, but which is tolerant to the positioning of the tandem repeats and to mismatches. Only after enrichment by this method are the fragments cloned, resulting in a preselected library in which a significant proportion of clones comprise simple tandem repeats. This allows the rapid screening for and identification of usefully polymorphic loci by the simple examination and comparison of loci in different, possibly unrelated, individuals. Since the selected clones contain short inserts, the effort necessary to identify and sequence the region of the tandem array is also reduced. The use of long tandemly repeated hybridisation targets in the present invention for hybridisation screening for minisatellite clones allows the isolation of relatively long arrays and tolerates mismatched variant repeat arrays, allowing the identification of a wide range of minisatellite clones, something which has been hitherto impossible to achieve.

In a first aspect of the present invention there is provided a method for the identification from DNA of a fragment comprising a simple tandem repeat locus comprising the steps of:

i) contacting a DNA library with at least one hybridisation probe so as to identify a population of DNA fragments enriched for simple tandem repeats;

ii) isolating and cloning said population; and

iii) screening of the resulting DNA library so as to identify an individual fragment comprising a simple tandem repeat locus.

The DNA library may be a genomic DNA library; the genomic DNA library may be any convenient population of DNA fragments such as human DNA, DNA from non-human species or subgenomic DNA libraries such as those generated by PCR from flow sorted chromosomes (see Telenius, H., *et al.* (1992). *Genomics* 13: 718-725). The genomic DNA library may be obtained by restriction digestion of genomic DNA.

The average fragment size within the DNA library may be less than 1.5 kilobases and may be less than about one kilobase. The fragment size may be from about 400bp to about 1000bp.

The hybridisation probe or set of probes may be immobilised on a solid phase such as a nylon membrane and may identify a particular class of simple tandem repeats. Such classes may include dimeric, trimeric, tetrameric, pentameric and hexameric tandem repeats such as trimeric or tetrameric repeats. Particular oligonucleotide probes for use in the present invention may include oligonucleotide probes comprising a tandemly repeated region of greater than 200bp. The probe may comprise repeats having at least 70%, such as 80% or 90%, similarity to a given repeat sequence. The hybridisation probe may be a set of probes comprising mixed trimeric or tetrameric repeat DNA.

The population of DNA fragments enriched for simple tandem repeats may be amplified prior to cloning and this may be effected by PCR amplification. Universal linker sequences may be ligated to the ends of individual fragments, possibly prior to the enrichment

procedure, and linker sequence specific primers may then be used to amplify the enriched population. Linker sequences may then be removed, for example by restriction digestion, prior to cloning.

According to the present invention there is also provided a method for the identification from genomic DNA of a fragment comprising a simple tandem repeat locus comprising the steps of:

i) ligating universal linker sequences to the ends of fragments comprised in a genomic DNA library so as to form a library for PCR amplification;

ii) contacting said PCR library with at least one hybridisation probe so as to identify a population of library fragments enriched for simple tandem repeats;

iii) separating and amplifying said population by PCR; and

iv) cloning and screening the resulting amplification products so as to isolate an individual fragment comprising a simple tandem repeat locus.

Cloning may be effected using any convenient cloning procedure and vector (for example pBluescriptII (Stratagene)) such as those described by Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory Press.

Screening may be effected using any convenient hybridisation probe or set of probes comprising simple tandem repeat sequences. These may be the same as those disclosed above in respect of the enrichment procedure. Individual clones comprising simple tandem repeat loci may be analysed using conventional techniques to determine for example specific sequence information.

By "simple tandem repeat locus" is meant a tandemly repeated region having a periodicity of up to eight bases, for example up to six bases, such as up to five, four or three

bases. Particular simple tandem repeat loci may have a periodicity of up to four or up to three bases.

The method of the present invention has been used to identify a number of simple tandem repeat loci as disclosed in Table 2, together with corresponding flanking primer sequences disclosed in Table 3 and hybridisation probes which specifically identify such loci.

Therefore, according to the present invention there are also provided simple tandem repeat loci for use in a method of treatment or diagnosis of the human or animal body characterised in that they may be amplified at least in part by PCR using any pair of primers as disclosed in Table 2.

The simple tandem repeats may comprise at least the sequence of at least any one of sequences 1-47. Where a pair of sequences are indicated (see Table 4 and sequences 1-47), the first part of the sequence may be separated from the second part of the sequence by an intervening sequence. This intervening sequence may comprise the repeat block of the simple tandem repeat.

The simple tandem repeats may be polymorphic. Many of the STR loci so identified have been shown to have unexpectedly high polymorphism. Therefore, they may have a heterozygosity of at least 80%; they may have a heterozygosity of at least 85%; they may have a heterozygosity of at least 90%.

The present invention also provides amplification primers specific to the aforesaid simple tandem repeats for use in a method of treatment or diagnosis of the human or animal body; the method of amplification may be PCR.

The present invention also provides probes specific to at least part of the

aforesaid simple tandem repeats for use in a method of treatment or diagnosis of the human or animal body.

According to further aspects of the present invention there are also provided methods of genetic characterisation wherein sample DNA is characterised by reference to at least one of the aforesaid loci, primer sequences and probes. The method of genetic characterisation may comprise either the use of at least one hybridisation probe or it may comprise the use of polymerase chain reaction (PCR) primers specific to at least one of the aforementioned loci in order to amplify selectively the simple tandem repeat locus. The PCR primers may comprise at least one of the primers and probes of the present invention. The method of genetic characterisation may be used in genetic mapping studies such as linkage studies, and may be used in the genetic analysis and diagnosis of inherited or acquired disease alleles.

Such techniques of genetic characterisation may allow the generation of individual 'identities' specific for one or more polymorphic loci, possibly those of the present invention. The generation of such individual 'identities' may be used to identify and characterise family relationships and may be used for e.g. forensic testing and in any technique which uses simple tandem repeats and their polymorphisms, such techniques possibly identifying, for example, inherited diseases and their causes.

Throughout the present application, the standard IUPAC nucleotide representation procedures are used. It should be noted that in these, R = A or G; Y = T or C; K = G or T; S = G or C; W = A or T; N = any base.

The invention will be further apparent from, but not limited to, the following description and examples with reference to the several accompanying figures and tables.

Of the figures,

Figure 1 shows a schematic summary of enrichment by filter hybridisation.

Figure 2 shows an enrichment for tandem repeats after filter hybridisation. Three replicate filters (A, B and C) bearing DNA from over 1000 clones from an enriched library were screened by hybridisation using a mixed triplet probe (left panel) or mixed tetramers (right panel).

Figure 3 shows examples of genotyping at polymorphic tetranucleotide repeat arrays using ^{33}P end-labelled primers and denaturing polyacrylamide gels; ten unrelated individuals were typed at wg1e12 (D7S822) and wg1c4 (D8S580). Estimated fragment sizes (nt) are indicated.

Figure 4 shows the number of alleles observed at the 46 tandem arrays identified in Table 1 investigated shown plotted against total array length. Triangular symbols denote triplet repeat arrays and squares denote tetrameric arrays; loci found to be monomorphic (one allele only) have been included in the analysis.

Of the Tables,

Table 1 shows characterisation by sequence analysis of 54 positively-hybridising clones from the repeat-enriched library, resulting in the first 24 polymorphic loci of Table 2

Table 2 shows properties of 24 polymorphic loci characterised in an initial study, together with thirteen subsequently identified loci. The number of alleles and heterozygosity

levels shown were those observed in the analysis of 20 or more unrelated individuals from the CEPH pedigrees. EMBL is the European Molecular Biology Laboratory data bank; GDB is the Genome Database

Table 3 shows PCR primer sequences and annealing temperatures for the polymorphic loci described; primers shown marked with an asterisk were end-labelled for genotyping on denaturing polyacrylamide gels.

Table 4 shows the simple tandem repeats and their sequence number or numbers. Where more than one number is given, the first part of the sequence may be separated from the second part of the sequence by an intervening sequence.

Results

During the characterisation of sequences isolated during an initial study (see Table 1), a total of 54 positively hybridising clones were analysed, giving 46 different sequences containing tandem repeat arrays (27 tetrameric; 19 triplet). These sequences were then used to design PCR primers (Table 1). Interestingly, two of these sequences showed near-perfect matches with sequences in the Genbank sequence database.

Of these sequences containing tandem repeat arrays, further characterisation revealed that 24 of them showed length polymorphisms when tested in 4 unrelated individuals. These polymorphic loci are shown in Table 2 together with 13 subsequently identified loci.

At these 37 new polymorphic loci, heterozygosity levels range from 9% to 95% (Table 2). The simplest factor useful in the prediction of variability was found to be the nature of the repeat block; tetramer repeat arrays not only showed more frequent polymorphism than triplets (18/27 v 6/19 in the initial study), but the average heterozygosity of those loci which

were polymorphic was also higher (75% v 34% in the initial study). The locus represented by clone wg1c3 contains a modestly variable triplet repeat (GGC)n array. However, the non-repetitive region of the sequence determined shows 98% similarity over 148 bases with the published sequence of the human cDNA for translation initiation factor 4D (eIF4D) (see Smit-McBride, *et al.* (1989). *J. Biol. Chem.* 264: 1578-1583). The region of near-identity begins abruptly at position 330 of wg1c3, corresponding to position 22 of the eIF4D cDNA, and suggests that the fragment isolated in clone wg1c3 may span an intron-exon boundary from the human eIF4D gene, and thus placing the (GGC)n repeat array within the preceding intron.

Although identified as containing tandem repeats by hybridisation, 5 of the 54 clones initially examined (about 10%) did not contain a recognisable tandem array. We assume that these clones might contain short imperfect arrays at some distance into the initial sequence analysis; where long (>8 repeat) arrays of near-perfect repeats were present they could easily be identified on sequencing autoradiographs, even when some distance into the clone. Alternatively, they might represent sequences rescued because of cross-hybridisation to a (non-repeated) part of one of the sequences contributing to the membrane-bound "target" DNA.

Hence the method of the present invention is a rapid and efficient method for isolating simple sequence loci, in this case exemplified by tri- and tetrameric repeat loci from the human genome.

Methods

The methods described below were used for the hybridisation selection of simple repeat loci, the characterisation of isolated sequences and the characterisation of these novel polymorphic simple sequence loci.

Hybridisation Selection of Simple Repeat Loci

Figure 1 shows the general strategy for the hybridisation selection of simple repeat loci. More specifically, human *Mbo*I fragments (400-1000bp) were ligated to (SAULA/SAULB) linkers to give a "whole-genome" PCR library (see Kinzler, KW. and Vogelstein, B. (1989). *Nucleic Acids Res.* 17: 3645-3653) from which tandem repeat-containing fragments were selected and reamplified. This population of molecules was denatured and incubated with two small nylon filters, one bearing mixed trimeric repeat DNA, the other bearing mixed tetrameric repeat arrays as described below. After hybridisation overnight at 65°C, fragments hybridising to each filter were recovered and reamplified using the linker primer SAULA. The reamplified fractions were compared with the input DNA by Southern blotting and probing with the pooled triplet or tetramer sequences used for selection, and were shown to be highly enriched for the respective sequences. Dot-blot analysis of serial dilutions showed that the enrichment was at least 50-100 fold.

DNA fragments from the reamplified, enriched DNA fractions were digested with *Mbo*I (to remove linkers) and cloned into pBluescriptII vectors. Clones from the resulting library, containing both triplet and tetramer selected sequences, were picked into ordered array and screened by hybridisation to check the frequency of repeat-containing clones (see Figure 2 for example). The probe mixtures originally used for filter hybridisation enrichment were also used as the hybridisation probes at this stage, and approximately 30% of clones initially studied gave positive hybridisation signals.

Characterisation of isolated sequences

The positively-hybridising clones were analysed for the initial characterisation of the enriched fragments. For each clone a first round of sequence analysis was performed using single stranded templates. In clones where the distal portion of unique sequence DNA

could not be determined directly, further sequence information was derived, either from the other end of the insert (using double-stranded plasmid template) or by using a specific primer proximal to the array to extend the region already sequenced. Where a specific primer was used for sequence extension, this primer was subsequently used as an amplimer for PCR.

Novel polymorphic simple sequence loci

Specific products were amplified by PCR from all of the 46 loci isolated. At most loci, specific amplification products could be satisfactorily resolved on agarose (NuSieve (RTM)) gels. At loci derived from expanded polyadenylate tracts, or where finer resolution was required, an end-labelled primer was used, and fragments resolved on denaturing polyacrylamide gels as described above; examples of tetranucleotide loci typed by this method are shown in Figure 3. The polymorphic loci identified are shown in Table 2.

The loci have been initially mapped using a combination of somatic cell hybrid analysis and linkage in a small number of CEPH pedigrees as described for minisatellite loci (see Armour, J.A.L., *et al.* (1990). *Genomics* 8: 501-512; and Armour, J.A.L., *et al.* (1992). *Human Mol. Genet.* 1: 319-323). At many loci, the placement of recombinations could be inferred from linkage analysis, and thus a tentative interval containing the locus could be defined even from the analysis of only two or three pedigrees. These putative placements determined by linkage analysis (Table 2) were made using the NIH/CEPH maps of the human chromosomes (see NIH/CEPH Collaborative Mapping Group (1992). *Science* 258: 148-162) as a framework. In all cases a LOD score of 3 or more was required as evidence of linkage.

General methods

The following general methods were used in the hybridisation selection of simple tandem repeat loci and the characterisation of isolated sequences.

General PCR conditions

PCR was carried out using the buffer described by Jeffreys, A.J., *et al.* (1990). *Cell* 60: 473-485 and 0.05U/µl Taq polymerase; unlabelled primers were added to a final concentration of 1µM. Cycling conditions were: for the whole genome PCR library, 95°C 1 min/67°C 1 min/70°C 2 min; for amplifying simple repeat loci using locus-specific primers, 95°C 1 min/ T°C 1 min/ 70°C 1 min; the annealing temperature T used for each polymorphic locus is shown in Table 3. These annealing temperatures gave good results in the PCR buffer used. In other buffer systems different temperatures may improve genotyping.

For amplification to levels visible after ethidium staining, 100ng genomic DNA was amplified for 32 cycles. Amplified fragments were resolved by electrophoresis in NuSieve (RTM) gels (FMC BioProducts, 2.5 - 4.5% according to fragment size) in 0.5x TBE buffer, and DNA detected by ethidium staining. Loci derived from expanded (Alu) polyadenylate tracts were most clearly detected after end-labelling of the non-Alu PCR primer and autoradiography of dried polyacrylamide gels; at other loci polymorphism could be detected on ethidium stained agarose gels, but the use of polyacrylamide provided added resolution of closely-spaced alleles. One primer was end-labelled (1.5pmol primer per subsequent PCR reaction) using [γ-^{33}P]ATP and T4- polynucleotide kinase. This labelled primer and 10pmol unlabelled primer were then used with 0.05U/µl Taq polymerase in a 10µl PCR. Table 3 shows the primer used for end-labelling marked with an asterisk. In general, 18 cycles were sufficient to give clean typing from 100ng genomic DNA; details of PCR conditions and primer sequences for the first 24 polymorphic loci can be found in GDB.

Whole genomic PCR library construction

SAU linkers were made by annealing equimolar amounts of SAULA (5'GCGGTACCCGGGAAGCTTGG3') and 5' phosphorylated SAULB

(5'GATCCCAAGCTTCCCGGGTACCGC3') as described by Royle, N.J., *et al.* (1992). *Proc. R. Soc. Lond.* B 247: 57-61. Human genomic DNA pooled from 20 unrelated individuals was digested with *Mbo*I and a 400-1000bp fraction size-selected after agarose gel electrophoresis. 200ng of DNA from this fraction were ligated with 2µg SAU linkers, a linker:fragment molar ratio of about 250:1. After a further round of size-selection to remove linker dimers, the library was amplified using SAULA primer to give products in the 400-1000bp range.

Tandem repeat "target" sequences

Both naturally occurring and synthetic sequences were used as target sequences in hybridisation selection. Where cloned or amplified loci were used as a source of tandemly repeated DNA, care was taken to use fragments which did not contain human dispersed repeat elements. The triplet repeat sequences ACC and AGG were selected using a DNA fragment from a cloned human locus, pMS633, containing about 2kb of interspersed AGG/TGG(=ACC) repeats. Tandem arrays of the other triplet sequences used (AGC, ACG, ATG, AGT and CCG) were synthesised as follows: 18mer oligonucleotides of each sequence and its complement were synthesised, phosphorylated, annealed and ligated into concatemers as described by Vergnaud, G. (1989). *Nucleic Acids Res.* 17: 7623-7630. Fragments larger than 200bp were size-fractionated from the ligated DNA and subjected to cycles of PCR in the absence of primers to selectively lengthen tandem (rather than inverted) arrays (see Collick, A. and Jeffreys, A.J. (1990). *Nucleic Acids Res.* 18: 625-629), and fragments of apparent size greater than 1000bp recovered from a 1% agarose gel. The triplet sequences AAT, AAG and AAC were not used in order to avoid heavy bias towards triplet repeats arising from retroposon tails (see Beckman, J.S. and Weber, J.L. (1992) *Genomics* 12: 627-631).

The self-complementary tetrameric repeat sequences CATG and CTAG were synthesised as 16mer oligonucleotides and assembled into long arrays by ligation and primer-free PCR. For other tetrameric sequences, cloned or amplified genomic fragments were

used. DNA from the tetramer repeat locus composed of ATGG repeats near the human myelin basic protein gene (see Boylan, K.B., *et al.* (1990). *Genomics* 6: 16-22) was amplified from human genomic DNA using the primers MBP1 (5'ACAAGGACCTCGTGAATTACAATC3') and MBP2 (5'ACAGGATTCACTCACATATTCCTG3'), to give fragments of about 1kb. DNA containing GGCA repeats was amplified from the mouse minisatellite clone p9.2 (see Gibbs, M., *et al.* (1993). *Genomics* 17: 121-128). A subcloned fragment from cosmid G2 (see Armour, J. *et al.* (1992). *Ann. Hum. Genet.* 56: 183) contains about 800bp of interspersed ACCC and ATCC repeats; the human minisatellite clone pMS630 (see Armour, J.A.L., *et al.* (1992). *Human Mol. Genet.* 1: 319-323) contains the octameric repeat (GGAGGGAA) and was thus used to select AGGG and AAGG tetramer repeats.

Hybridisation Selection

DNA fractions containing the different trimeric arrays were pooled, denatured by treatment with alkali (KOH, final concentration 150mM), neutralised by adding 0.25 volumes of 1M TRIS-HCl pH 4.8, and a total of 1µg spotted onto a small (3mm x 3mm) piece of nylon filter (Hybond-N^{fp}, Amersham). When dry, the filter was exposed to ultraviolet light to bind the DNA. A similar small filter was made using the pooled tetrameric repeat fragments. Since the filters were to be used to select different types of sequence from the same input DNA, they could be used together in the same hybridisation. The filters were prehybridised in 1ml phosphate/SDS buffer (see Church, G.M. and Gilbert, W. (1984). *Proc. Nat. Acad. Sci U.S.A.* 81: 1991-1995) at 65°C and transferred to 100µl of the same buffer at 65°C in an Eppendorf tube.

Input DNA was amplified from the whole genome PCR library and about 1µg denatured with alkali, neutralised, and added to the buffer containing the filters; the reaction was covered with paraffin oil and incubated overnight at 65°C. The filters were washed thoroughly in 0.2x SSC, 0.01% SDS at 65°C. After washing, the DNA bound to each filter was removed by treatment (at room temperature) with 50µl of 50mM KOH/0.01% SDS, followed by 50µl of

50mM TRIS-HCl (pH 7.5) / 0.01% SDS; the washings from each filter were pooled, and recovered DNA ethanol precipitated using primer SAULA (final concentration 1μM) as a carrier.

Hybridisation-selected DNA was reamplified, digested with *Mbo*I and cloned into pBluescriptII vectors (Stratagene). Clones were picked into ordered array for ease of screening and clones replicated from microtitre plates onto Nylon filters in groups of four as described by Brownstein, B.H., *et al.* (1989). *Science* 244: 1348-1351

Sequence determination and analysis

DNA sequence was initially determined from clones hybridising positively with tandem repeat probes using dideoxynucleotide chain termination with T7 DNA polymerase (see Tabor, S. and Richardson, C.C. (1987). *Proc. Nat. Acad. Sci. U.S.A.* 84: 4767-4771) on single stranded templates. Where this proved insufficient to determine sequence on both sides of a tandem array, additional sequence was determined either from the other end of the clone using double-stranded templates, or after extension of the sequence using a specific oligonucleotide primer; any specific primer used in sequence analysis was subsequently recruited as one of the amplimers in PCR.

The sequence determined was analysed using the suite of programs developed at the University of Wisconsin (Genetics Computer Group (1991). Program Manual for the GCG Package. Version 7. April 1991. 575 Science Drive, Madison, Wisconsin, USA 53711); updated sequence databases were searched using the BLAST (see Altschul. S.F., *et al.* (1990). *J. Mol. Biol.* 215: 403-410) network service at the NCBI.

It will be appreciated that it is not intended to limit the invention to the above example only, many variations, such as might readily occur to one skilled in the art, being

possible, without departing from the scope thereof as defined by the appended claims.

DNA Sequence Information

Sequence: 1, corresponds to wg0e7; Length 377

5' gatcaaattt attctctcct ttgcacactg gaagtgcaag taacatttct

51 tccttctcct gctcctcctc ctgataacaa tggtgatgat gatggtgatg

101 atggtggtgg tgatggtgat gttggtggtg atgatggtga tggtgatggt

151 ggtgatgatg gtggtggtga tggtggtggt gatggtggtg gtgatggggt

201 ggtgacggtg atgttgacgg tggtggtggt ggtggtgatg gagtggtgat

251 ggagtggtga tgatggtggt gatggtggtg atggcgataa caaacatata

301 ttaagacctt accatggctr ggcatggtgg ctgatrcctg taatcccagc

351 actttgggag gccgaggcgg gcagatc

Sequence: 2, corresponds to first part of wg1a2; Length 346

5' gatcattcgg aagaaagtgt ggaagcagca gcaaagagtg gaaaatgaaa

51 agagaaactc tggagaaggc aaggtgggca ggagcaggac tgtgccgcct

101 gcacccatgc aggctaggcg ttgtccaaca ctggggcacc cgtcactcag

151 attgagatga gggacaatga gaggagcctg gaggagagct ccacacaaat

201 aaagggagaa gcctatgcag gggctggaga ttccttctgt ggtgacagag

251 catggcatag ttagattcac agactnnnnn nagatcgaga gaatgatgcg

301 tgctcttctc atctctcaag cagcaatgca gggggaacat cagctg

Sequence:3 corresponds to second part of wg1a2; Length 57

5' ttgttttttt gatggagtct cactctgttg cccaggctgg agtgcagtgg

51 cgtgatc

Sequence: 4 corresponds to first part of wg1a3; Length 217

5' gatccatcca tccttcctcc ctccttcccc ctctttctac ctctttctcg

51 ctctctcttt cttccttcct tcctcccctc ctcccctcct cccctcctcc

101 cttcctcccg ccctccttcc cttttcccct ccccccttcct ctttctttct

151 ctttctttct ctttctttct tactttcttt ctttctttct ttctttcttt

201 ctttctttct ttctttc

Sequence: 5 corresponds to the second part of wg1a3; Length 43

5' gagtctcact ctgttgccca ggctggattg cagtggcagg atc

Sequence: 6 corresponds to wg1a9; Length 286

5' gatcagtttc tgactgctgg gcgggacaaa gcctcctgaa gttgctgcga

51 ggcacctccc cctgtgagca gagcttggta cagcccaaat agttttcagg

101 ttaagaaagc cagaatcttt gttcagccac actgactgaa cagactttta

151 gtggggttac ctggctaaca gcagcagcgg caacggcagc agcagcagca

201 gcagcagcag cagcagcagc agggctcctg ggataactca ggtgagtaga

251 gagggaattc gcaaacttac cctggagttt tatttc

Sequence: 7 corresponds to wg1c3; Length 457

5' gatccaatgg ctctttagtc agggtgttat gtcctgaaaa taggtgacaa

51 ctgcaaacca tcctctggtg tccagagact ttaacaaggt ttgtttcaca

101 gagactgagg gcagaaaaaa ggaaatggcc taaaaaggtg ggtttgctgt

151 gttgcctcac actacttgat tcatggttct gattctaaaa atctcacttg

201 atacttgatt tcatatgaaa gacgtgtaaa atgcctgggt agaggcggcg

251 gcggcggcgg cggcgggctc ggaggcagcg gttgggctcg cggcgagcgg

301 acggggtcga gtcagtgccg tttgcgccgg ttggaatcga agcctcttaa

351 aatggcagat gatttggact tcgagacagg agatgcaggg gcctcagcca

401 ccttcccaat gcagtgctca gcattacgta agaatggctt tgtggtgctt

451 aaaggct

Sequence: 8 corresponds to wg1c4; Length 370

5' gatcacacca ttgcactcca gcttgggcaa cagagtgaga ctccatctca

51 acaaaaaaaa gaaagaaaag aaaagaaaga aaagaaaaga aaagaaaaga

101 aagagaaaga aagagaaaaa gaaagaagga aggaaggaag gaaggaagga

151 aggaaggaag gaagggagga aggaaggaag gaaagcaaga aagaaagaaa

201 gaaagaaaga aagagaaaga aagaaactat ccaaaccaat ctgatagagc

251 tgaaaaactt actacaagaa tttcataata caatcagaag tattaacaac

301 aaaatgcacc aagctaagga aagaatctca gaactagaag acccagttct

351 ttgaatctat tcagacagac

Sequence: 9 corresponds to wg1c5; Length 367

5' gatctcaata aacattgata ctggagggat gaaatgaagg aaggatggat

51 agaaggctat aaggatgggt ggatggataa atggatggat ggatggatgg

101 atggatggat ggatggatag atggatggat ggaaaaatgg atagatggat

151 gggtggatgg atgaatggat ggatggatgg atggatggat ggatgggtgg

201 atggatgaat atattgggtg gatggatgga aggaaggaag gaaggaagga

251 aggaaggaag gaaggaagga aggatggtag aagaaaggta gtaccagtat

301 gctttagctc atgcaggcaa acagatgatg ggcagaggga agcatggtgg

351 ctgattacag gaggatc

Sequence: 10 corresponds to wg1d1; Length 434

5' gatcctcttg cctaggcctc ccaaagtgct gggattacag gcaagagcca

51 ccacgtcccg cctctaattt ctctcctctt ctctcctctc ctttcctttc

101 ctctcctctc ctttcctttc ctcttctctt cttgtttttt tcttttctnc

151 cctccctccc tccctccctc cctccctccc tccttccttc cttccttcct

201 tccttccttc cttccttcct tccttccttt ttgagacaga gttttattct

251 gtcacccaga cctgagtgca atgggcacaa ttttggctca ctgcaacctc

301 catctccccg gttcaagtga ttcttctgcc ttagcctccc gaatagctgg

351 aactacaggc acctgccacc atgccccagc taatttttg tattctcagt

401 agagatgggg tttaccatgt tggccaggct gctc

Sequence: 11 corresponds to wg1d5; Length 325

5' gatcgcgcca ctgtactcca gcctgggcga cagagcgaga ctctgtctaa

51 aaaaaaagaa aaaagaaaaa aaagaaagaa tgagagaaag agagaaggaa

101 ggaaggaagg aaaaggaaag agagagagga aggaaggaag gaaggaagga

151 aggaaggaag gaaggaagga aggaagggag ggagggaggg agggagggag

201 ggagggaggg agggaaaggc agggagaaag ttctgggagc tagggagtgc

251 ccggggtggg gagctccaag aacaagcccc agggagctgt aacaaagact

301 ttgtcacagc tagcctgaag ctagc

Sequence: 12 corresponds to wg1d6; Length 263

5' gatcccacct gccatacggt gggatttcta ggactataca aatgacagaa

51 gggtagtaag aggaagactg tgttgcttaa tgaggtttcc agaaattggt

101 aatgatattt gtaattccaa atcctactac aaggaactgt ggctacaata

151 ttgatgctgc tgctgctgct gctgctaatt tgatgaagta ggctaatccg

201 catggctaca tctctgtatt agtccattct cgcgctgcta taaagaaact

251 acctgtgact ggg

Sequence: 13 corresponds to wg1d10; Length 160

5' gatcctgttc atggtacaaa gctttcccta gcagcctgcc ctccctagcc

51 tgcttacctt gagnngagag gaagctgaag tagcagcagc agcagcagca

101 gcagcagagt tnccagaaag tgacccccctc ccctgaacac agcaggaagc

151 agcagtccaa

Sequence: 14 corresponds to the first part of wg1d11; Length 238

5' gatcatttca gtctgcacaa gaatgcttgg ccttttaatt ccaacttcac

51 agttgagaaa actgatactc aaggcaaaga atcttctcag tagtcagagt

101 caataactgc aggaactaag actggaaccc aagttttctg cctggtatgt

151 tgggcctaga agggaactgc tattcctatc tctccatctt tccttccatc

201 tttccttcct tccttccttc cttaatcctt ccttcctt

Sequence: 15 corresponds to second part of wg1d11; Langth 120

5' ttcctctctc actgtctccc tcnctctctg tctccctcct tcctttcttc

51 accttctttc tacttttta agaaacaagg tctggctttg tcacccaggc

101 tggagtgcag tggcgtgatc

Sequence: 16 corresponds to wg1e1; Length 445

5' gatcttgaga cagggtcatc ctggattact ggagtgtgcc ctaaatccat

51 tgacaagtgt ccttaggaga gacgcagagt ggaggcacac agtgggagga

101 cgaggccact tgaagactga ggccgggatt gcagcgatgc agccacaacc

151 caggaaagtc cggggccacc agcggctgga aaaggcaagg gaggggtctt

201 ctggctcttc aacaataaga gagtaaattt ctggtgtttt aagccacctg

251 gtttgtggtg ctttttcctt ccctccttcc ttccttcctt ccttccttcc

301 ttccttcctt ccttccttcc ttccttcccc ccctccctcc ctccctctct

351 ccctccctcc ctcccttcct tcctcttctt tttctctccc tctctccttt

401 ttttcttttt tttggtggag tcttgctgtg tcgcccaggc tggag

Sequence: 17 corresponds to wg1e4; Length 591

5' gatcccaaaa tactggcctc tcatagtgat agatttaaaa gattgcttct

51 ttaccattcc tttagctacc caagattatg aaaaatttgc ttttactgtt

101 ccttctataa ataacaaaga accagtggac agataccatt ggaaagtact

151 gccacaaggc atgctaaata gcccgactat ttgtcaaact tatatcggaa

201 aagttatgaa gccaattaaa gaacaatttt acaaatgtta tattatccat

251 tacatggata attttatttg cagctgaaac taaagaggaa ttaatgctat

301 gctacaaaca actggaaaag gctgtgactg cagcgggatt aatcaatcat

351 agccctgata aaatccaaac ttctactccc tttcagtatt taggaatgaa

401 agcagaataa agtactatca agcttcaaaa ggttcaaatt agaagagatg

451 atttaaaaac tctaaatggc cggcctgcct tccttccttc cctccctccc

501 tccctccctc cctcccttcc ttccttcctc cctccctctc tctttcgacg

551 gtctccctct gttgccgagg ctggactgta ctgccatgat c

Sequence: 18 corresponds to wg1e7; Length 485

5' gatcacttga ggccaggagt tcaagaccag cctgagcaac ctagtgaaac

51 cccgtttcta caaaaaataa aaatttaaga aatagctgga tgcagaggca

101 tctgcctgta gtcccagcta cccaggaggt tgaggaagga gaatcacttg

151 agcccagaag cttgaggttg tagtaaggaa tgttcatgcc actgcactgc

201 aacatgggtg acagtgcaag tttctgcctc aaaaggaagg aaggaaggaa

251 ggaaggaagg aaggaaggaa ggaaggaagg aaggcaggca agaagaaaag

301 aaggcaggga gagacggagg gaaagacaga aaagaaagaa aacctataaa

351 aaagtataat cctgtgagtc cacagatgag acagagaaaa atctggaaag

401 gattttaaaa taagtatgct taaattcttc aaagagacat agaaaggaat

451 agaacccaca aaataagaat ggaaatattc gaaaa

Sequence: 19 corresponds to wg1e12; Length 597

5' gatcttatga cattttccca ggacaccaag atataaaacc ccaaccaaca

51 ttgctactgc taaagtaaac ttttgcctgg cttgccaaga tttttggcca

101 agaaatgaga tttcctgagg gtggcattcc ctctgcacta ccaaagtctc

151 cttctgagac tttttggtca gcttatgaag cttctcaagg caagtgtctg

201 gttagcatct ccctccctcc actctggaaa tcttaaagct gaaagaatga

251 atgaatgaat ggatgaatga atgaatgaat gagaagacag agagagagaa

301 ggaaggaagg aaggaaggaa ggaaggaagg aaggaaggaa agaaagaaaa

351 gaaaagaaag aagagagaaa gagagaaaga aagaaagaga gaaagagaga

401 gagagagaga gagaaagaga gagaggaaga gaagaagtcc tcttaaaaaa

451 tagcctgaga aactgggcta tgttggcttt ttttttttc tgtcagtagg

501 aaatatttat tcaacctcac tgctaaaaaa aaaccaaaac aaacaaacaa

551 aaaaacctaa taatttcagg aaagctgctg tttctcgtgt tctgatc

Sequence: 20 corresponds to wg1f2; Length 350

5' gatcacgcca ctgcactcca gcctgggtga cagtgtgaga ccctgtcaag

51 gaacgaacga aggaaggaag gaaggaagga aggaaggaag gaaggaagga

101 aggaaggaag gaaggaagga aggagggaag gaaggaaaga aggcaggcag

151 gcaggcaggc aggaaggaag gaaggaaaga aggaaggaag gaaagaagga

201 aggaaggaaa gaaggaagga aggtaggaag gtaggaagga aggaaggaag

251 gaaagaagga aggaaggaag gcagtcaggg agnaaggaag gaaggcaggc

301 aggcaggcag gcaggcaggc aggcttgcaa atgtagttaa gttaaagatc

Sequence: 21 corresponds to wg1f4; Length 283

5' gatcatgcgg gcagctttgg ggtatttcag acggtgtggg gagcatggtc

51 tgaatgtgcc ttgctccggc agcagcatgc agtagtggca gtggtactta

101 gggcatgtga gagcaccctg cctctcctat ccctgaccca gcagcatgca

151 gtagcggcag tggtacttag ggcatgtgag agcacccagc ctctcctatc

201 cctgacccag cagctggcag cagcagcagc agcagcagca gcagcagccg

251 cctcagggca ggaggcagag ccttcaggcg tgg

Sequence: 22 corresponds to wg1g5; Length 494

5' gatcactgca ctccagcctg ggtgacagaa taagacgaaa gagagaaaga

51 gagagggaaa gaaagaaaga gagagagaga gagagagaga gagagagaaa

101 gaaagaaaga aagaaagaaa gaaagaagaa agcaagaagg aaggaaggaa

151 ggaaagaaag cagcagaaaa agaggaaggg agggaggaag gaaggaagga

201 aggaagggag ggagggaagg aaggaaggaa ggaaggaagg aaggaaggaa

251 ggaaggaagg aaagagagag agagaaagaa aatannnnnn nnnnnnaact

301 ccnnnaaacc cacaattcag acacacagct cacacacagg tctccagcat

351 agacatattt atacatccat ttactcaaac actcacaata caatcacata

401 aaacaggcag acagttcaca tgccaacaca ctcttgcaca gacacgcaaa

451 cagaagcatg gaatttgtac agagcacgct cacagtgtct gatc

Sequence: 23 corresponds to wg1g9; Length 301

5' gatcgtgcca ctgtacccca gcctgggcta cagagcgaga ctccatctca

51 aaaaaaaaaa agaaagaaag aaagagagaa agagagagag agagatgaaa

101 gaaagagaga gagggaaaga aaggaaggaa ggaaggaagg aaggaaggaa

151 ggaaggaagg caggcaggca ggcaggcagg caggcaggca ggcaggcagg

201 cggacagcaa gaagacaccg ttttgccatg aggttagaca cgcggacagg

251 cacagagcag acgcacgtgc accatgctat catggcagga caggttcaca

301 t

Sequence: 24 corresponds to wg1h10; Length 538

5' gatcatcaaa atacaattat agaaatattt ataagcagca ttattcataa

51 tcgccaaaaa ctggaagcag tgcgatggca aaatagatgc ataaatggtg

101 ataagtataa gaggggaaga aagaatgaaa gaaagaatgg aaggaaggaa

151 ggaaggaggg aaggaaggaa ggagggaagg aaggagggaa ggaaggaggg

201 aaggagggaa ggaaggaggg aaggaaggaa gggaggaagg gagggaggga

251 gggagggagg aagggaggga gggagggagg gagggaggga aaggactaga

301 gggtggaaga tagggagaga aacaagtaaa taagctagct ctttcctaga

351 aaataatttc accaacgttt ctgtgacatt caagaaaaca actgggactt

401 ggaaacaatt aaaaataaat aaacaaaagt atgccactag actctaaagt

451 cagtggtgtg ggaagcagag gttatcagtg ttcagaggag agaagactcc

501 cacagaatag ggctgtcagg aatgagctca gggaggaa

Sequence: 25 corresponds to wg2a5; Length 421

5' gatcacctga ggtcaggagt ttgagaccag cctggacaac atgacaaaac

51 ccctatctaa aaaaaaagaa atagctaggc atggtggtgt gcacctgtag

101 tcccagctac ttgggaggct gaggcagaga atcacttgaa cccgggaggc

151 agaggttgca gtgagccgag gaggcgccac ttcactccag cctctgtctc

201 caaagaaaga aaggaaaaga aagaagggaa gaaagagaga gggaggaaag

251 aaggggagga aggaaggaag gaaggaagga aggaaggaag ggagggaggg

301 agggagggag ggagggaggg agggagggag ggagggaaaa agaaaagagg

351 tgagcacacg gttacattga ggaaaacaaa gatgaaactt cacatcacat

401 tccaacaagt cacagcttga t

Sequence: 26 corresponds to wg2b3; Length 446

5' gatctcaggt gacccaccag cctcagcctc ccaaagtgct gggattacag

51 gcctgtgcca ctgcacccag ccatctgttc agtactttca ttataagaga

101 gaaaggagga gagggaaggg aagggaggga aggggagggg aagggaatca

151 atgggaaagg agggtcaaga aggagaagga gagaaggaag gaagggaggg

201 aggaagagag gaaggaagga aggaaagaag gagggaagga aggaaggaaa

251 gaaggaggga aggaaggaag gagggaggag ggaaggaagg aaggaaggaa

301 agaaggaagg aaggaaggaa agagggaaga aaggaaggaa ggaaagaaag

351 aagggaggaa gggagggagg gagggaggga atgagtggna gaagccaagt

401 ctgcagttgg gaaatcatgg gacgtgctgg cttttcctct ctgatc

Sequence: 27 corresponds to wg2c9; Length 287

5' gatcacttga gcccaggagt tcaaggctgt ggtaaactat aatcacacta

51 ctgcactcca gcctgggtga cagagaaaga ccctgtctca aaaaaggaag

101 gaaagaagga aggaaggaag gaaggaagga gggagggagg gagggaggaa

151 gggagggagg gagggaagga gggaaggaag gaaggaagga aggaaggaag

201 gaaggaaata gcagctctga gcttagaaaa aggagtctat ttctaagtgg

251 gagatgggga gaaggaggga actggggagg tgaggaa

Sequence: 28 corresponds to wg0c5; Length 252

5' gatcattagg ttgaaaaaga gctaaaagat gaaaccgatt ggcactggtg

51 tgtggtggtg gtggaggagg tggtggtggc ggcggcggtg gtggtggtgg

101 tggtggtggt ggtggcggtg gtggtggtag gaattactca agttactgga

151 aacatgctgg tatcttttt tagtttaggt agtaaacctg gtaatgaaca

201 ctaagtcaaa caacaaatac taatttccat ctcatgcaca aatgatatga

251 aa

Sequence: 29 corresponds to wg0f4; Length 329

5' gatcagacac tctaaagtca cattccttta gaggaactgg acaatcaaat

51 tttgatggtg ttctaatggt ttgtaaggca acaaaacaca aaactttgtg

101 gtggtggtgg tggtggtggt ggtggtggtg gtggtatctt ccatcacttg

151 ccaagggctt agcctggacc tgcacactca ctatctcctt gaccatttgc

201 accatcacca ggagggaggc actaggtccc ccgttctcac tgttataaat

251 aacaaacagg tctccaaggg gtgagtaact ttctcgtgga cacacagagg

301 caggtctagg atttgaaccc agtttgtct

Sequence: 30 corresponds to wg0f5; Length 276

5' gatctctcta ggtcctattc tctttcaacc ctctagggaa ctcaggaaac

51 attgggctat tgtccataat gtggtgatgg tggtggtgat ggtggtggtg

101 gtggcggtga tggcagcggc agtggtgatg gcgatggcgg cagcggcggt

151 ggtggtggtg gtgtcacccg aggctgcctt ggtccagcca gcacgcagcc

201 ttctctattc attctctctt gtgtggaccc gtgggggaat tctatgagtc

251 ttgccacttc anggctccac tcagaa

Sequence: 31 corresponds to the first part of wg2e7; Length 185

5' GATCATAGAG CAGGTCACCA GGATGAAGAC TGCATGAAGG CAAGGGCTTT

51 GATGTACTCA TTGTCCTGGC CCCGGCATGG AGGTGCTGGA AGGCAAGAGG

101 GAGGAGGAGG GAGGCAGAGA TGGAAGGATG AAGGAGAAGA AGGAAGGAAG

151 GAAGGAAGGG AAGGAGGGAG GGACAGAGGG AGGAT

Sequence: 32 corresponds to the second part of wg2e7; Length 22 (bases 193-214)

5' GGAAAGTTTT TTTAAAAAGA TC

Sequence: 33 corresponds to the first part of wg2f7; Length 140

5' GATCTACATG CATAGTTTAT TTTTTATGTT CTTTTATGTT TGTTAATATG

51 TAAATATATT TGTGATATAT TATTAAGTNA GAATATCAAC NGCCTTCCTT

101 CTTTCCNNCC CTCCCCACTT CCCTNCCTTN CCTTCCCAGC

Sequence: 34 corresponds to the second part of wg2f7; Length 211 (bases 144-354)

5' TCTGACAAGG TCTGTCTCTG TCACCTAGGC TAGAATGCAG TGGTGNAATC

51 AATAGCTCAC TGCAGCCTTG ACCTTATGGA CTCAAGTAAT CCTCCTACCT

101 CAGNNTCCNN ACAGNNGGGA CCTCAGGTGC ATACCACGCT CTGCTAATTT

151 ATAGAGATGG AGTCTTACCA TTTTGCCTAA GATGGTCTCC AACTCCCGGG

201 TTCAAGTGAT C

Sequence: 35 corresponds to wg2f10; Length 374

5' GATCTTGGCT GGGTCAACAC TCCTTCCTGG GCTTCAGTTT CTCATCTAAG

51 AAGAGAGAGT TGGAGGATTG TGGTGGGGGG TTGGTCAGTG AAGGTAGGCA

101 TCCCAGGGTG GGTANCCATG AGGGTCTCTC TAGTCCTTTT TTCTTCTTCA

151 CCCTTACACT TATCCACCCA TCCAACCATC CATCCATCCA TCCATCCATC

201 CATCCATCCA TCCATTTTTT CTTTTTTCTT TTTTTCTTTT TTTGAGATGG

251 AGTCTTGCTC TGTTGCCCAG GCTGGAGTGC AGTGGCATGA TGTCAGCTCA

301 CTGCAACCTC TGCCTCCTGA GTTAGAGTGA TTTTCCTGCC TCAGCCTCCT

351 GAGTAGCTGG GACTATAGGC ACAC

Sequence: 36 corresponds to the first part of wg2g4; Length 106

5' GATCACCTGA GGGAGCTCAA GACCAGCCTG GCCAACATGA TGAAACCCCG

51 TATATACTAA AAAGTACAAA AAATCANNNG GGTGTGTGGT GGGANTGTAA

101 TNTTAG

Sequence: 37 corresponds to the second part of wg2g4; Length 124 (bases 114-237)

5' GAAAGAAAGA AAGAAAGAAA GAAAGCAAGC AAGCAAGCAA GCAAGCAAGC

51 AAGCAAGCAG GCAGGCAAGN NAGCGGCGTC ACGCCNGTAA TCCCAGCACT

101 TTGGGAGGCC GAGGCGGGCA GATC

Sequence: 38 corresponds to the first part of wg2g12; Length 213

5' GATCATTTCC CAGTACATAA GGACCTGTTT CTCTCCTGCT AACATTAACC

51 CTACTTGAGA CTTAGAGAAA GAGGCATCAC ACTTGAAAGT CTCCTGTGGG

101 TATAATGTCT ACTCTTTGTT TCATGAAAGG ATATCCTGGG GTGGTAGCTT

151 TTTGGTTTTC TTTCTCTCTT TCTCTCTTTC TTTCTTTCTT TCTTTCTTTT

201 CTTTCTTTCT TTC

Sequence: 39 corresponds to the second part of wg2g12; Length 67 (bases 224-288)

5' TTCCTTCCNT CTTTTTTGTG GATGGAGTTC TGCTCTGTCA CCCTGGCTGG

51 AGCGCAGTGG CACGATC

Sequence: 40 corresponds to the first part of wg2h11; Length 97

5' GATCGCACAC TGCACTCCAG CCTGGCAACA GAGGGAGACT TCATCAGAGA

51 CAGAGAGAGA CAGAAAGAGA GAGAGATAGA GAAAGGGAGG GAGGGAG

Sequence: 41 corresponds to the second part of wg2h11; Length 95 (bases 105-199)

5' AAGGAAGGAA AGAAGGAAGG AAGGAAGGAA GGAAAAAAGA AAAGAGAAAA

51 AAAAAGGAGA GAGGTTGAAA AAAACAACTA CCTTGTGGTC AGATC

Sequence: 42 corresponds to wg3a6; Length 278

5' GATCACTTAG CCTGGGAGGT TGAGGCTGCA GTGAGTCATG ATTTTGCCAC
51 TACTGCATTC CAGCCTGAGT GACAGAGCCA ACCTGTCTTG AAAGAAAGAA
101 AGAAAAGAAA GAAGGAAAGA AAGAAAGAAA GAAAGAAAGA GAGAAAGAAA
151 GAAAGAAGGA AAGAAAGGAA GGGAAAGAAA GAAAGGAGGG AGGGAAGGAG
201 GGAGGGAGGG AAGGAGGGAG GGAGGGAGGG AGTATAAGAT GTATCCCCTT
251 AGCAAATGTT TAAATACACA GTATAGTT

Sequence: 43 corresponds to the first part of wg3b10; Length 204

5' GATCAAAACT GAGAAGCGCA AAGACAAAGA GTGTGCTTGT TGAATACCAA
51 GTTGTATAGG CTGCAGAAGA GGAAGTGGTG GGACTGGAGT CTAGAGAGTC
101 TTGAACACCA GGTTTGGGAG TCTGGAGTTC ACTTGGTGAG TAACAATCTC
151 TGGCAGAGGA AGACTCCGTC TCAAAGAAAG AAAGAAAGAG AGAGAGAGAG
201 AGAG

Sequence: 44 corresponds to the second part of wg3b10; Length 85 (bases 212-296)

5' AAAGAAAGAA AGAAAGAAAG AAAGAAAGAA AGAAAGAGAG GAAAGAAAGA
51 AAGAAAAGAA AAAAAGGAAA GGAATGAAAG GGATC

Sequence: 45 corresponds to the first part of wg3f12; Length 140

5' GATCATGCTA CTGCACTCCT GCCTGGACGA CAGATTGAGA CCCCATCTCG
51 GAAGGAAGGA AGGAAGGAAG GGAGGGAGGA AGGAAGGAAG GAAGGGAGGG
101 AGGGAGGGAG GAAGGGAGGG AGGGAGGGAG GGAGGAAAAC

Sequence: 46 corresponds to the second part of wg3f12; Length 153 (bases 150-302)

5' ATAGAAAGTA AGAAAGAAAG GAAACAATTG TGTGATGCAC AGCTTTGTGC
51 AGTGAGGNTT TTTTTGCCTC CAAGGTTTTG GGACAAGAAG GCACACAGAG
101 AATTAAAGGA GTCCAGAGTT ACTTGCTGTC CTGATATAGA TCCACTAGTT

150 CTA

Sequence: 47 corresponds to wg3h2; Length 278

5' GATCTTTTTG GCTTTTTGGC ATAACATGGC TGGCAGAGCT CAAATTGTTT

51 TTATCAGCTT AGTTACCTCT ACCCAGTAGA AATACAACTG CTGAAATTGT

101 AATTAGGTCT TTTATATTCC TCTCCTTCCT CCCTCCCTCC CTCCCTCCCT

151 CCGTCCCTCC CTTCCTCCCT TCCTTCCTTC CTTTCTTCCC TACCCCCCTC

201 TCTTTCTTCT TTTTATTTCC TTGTTTATTT CTGTCTAGCA CTAGATTTCA

251 TGGGAGACAT AGACTAAGAT ATAAATTT

Table 1

	Triplet	Tetramer	Total
Repeats not seen	-	-	5
Loci represented twice	2	1	3
Distinct repeat arrays	19	27	46
Number polymorphic	6	18	24

Table 2

Clone	Major Repeat(s)	No. of Alleles	Heterozygosity (%)	Size Range	Chromosome	Interval	EMBL	GDB
wg0e7	AYC	4	50	265-330	12	-	X73968	D12S387
wg1a2	AARG	7	80	435-480	7	D7S369-D7S129	X74431/2	D7S808
wg1a3	AAAG	10	85	414-462	20	OXT-D20S27	X75550/1	D20S212
wg1a9	RRC	2	30	101-104	X	DXS178-DXS100	X73287	DXS1252
wg1c3	GGC	2	30	97-106	10	D10S7-D10S58	X76102	**(eIF4D?)**
wg1c4	AGGR	16	95	187-272	8	D8S85-D8S32	X74778	D8S580
wg1c5	AWGG	7	90	294-318	14	D14S53-D14S18	X73288	D14S299

wg1d1	AGGR	8	90	186-218	15	CYP1A1-D15S37	X75552	D15S233
wg1d5	AGGR	7	70	210-234	22	D22S257-PDGFB	X74779	D22S442
wg1d6	AGC	2	40	72-75	4	D4S174-D4S1	X73289	D4S1639
wg1d10	AGC	2	25	85-97	2	D2S44-D2S103	X74433	D2S419
wg1d11	AAGG	2	20	402-417	15	ACTC-D15S35	X75553/4	
wg1e1	AGGR	9	80	195-233	X	DXS297-qter	X75409	
wg1e4	AGGR	2	60	263-267	1	D1S81-D1S51	X74434	D1S525
wg1e7	AAGG	4	70	110-122	3	D3S13-D3S20	X74780	D3S1749

wg1e12	AAKG	9	75	266-298	7	D7S96-D7S129	X75555	D7S822
wg1f2	AAGG	13	90	280-502	16	HP-D16S43	X75556	D16S543
wg1f4	AGC	2	30	71-74	6	-	X73969	D6S483
wg1g5	AGRR	11	85	301-374	19	-	X74435	D19S428
wg1g9	AGGS	9	95	241-289	7	D7S81-D7S129	X73290	D7S809
wg1h10	RRGG	6	50	247-311	18	-	X74781	D18S555
wg2a5	ARGG	6	70	200-228	4	D4S175-D4S118	X73970	D4S1640
wg2b3	ARGG	2	50	360-380	20	D20S18-D20S17	X74436	D20S204
wg2c9	ARGG	7	95	187-247	1	D1S48-D1S68	X73291	D1S526
wg0c5	NCC	3	9	118-144	3	D3S13-D3S20	X73967	D3S1747
wg0f4	ACC	3	28	84-93	18	-	X73214	D18S496

wg0f5	RYC	2	53	208-217	2	D2S48-D2S34	X73215	D2S413
wg2e7		4	53	340-372	20			
wg2f7		3	45	254-262	17			
wg2f10		2	50	230-234	19			
wg2g4		>10	75	480-520	19			
wg2g12					2			
wg2h11		9	80	300-436	2			
wg3a6		12	90	227-275	11			
wg3b10		15	88	460-544	19			
wg3f12		5	71	240-280	11			
wg3h2		2	10	190-194	11			

The locus wg2g4 has at least 10 very closely spaced, large alleles. The values given for wg2g4 are a minimum estimate of the number of alleles andtheir size range, the figure for heterozygosity being for observed heterozygosity.

Table 3

Clone (locus)	Primer Names	Primer Pair	Primer Sequences (5' → 3')	EMBL	T°C
wg0e7 (D12S387)	wg0e7a wg0e7b	1	GCACACTGGA AGTGCAAGTA AC AGCCATGGTA AGTCTTAATA TATG	X73968	65
wg1a2 (D7S808)	wg1a2a* wg1a2b	2	TCTTCTCATC TCTCAAGCAG AGTGAGACTC CATCAAA	X74431/2	58
wg1a3 (D20S212)	wg1a3a* wg1a3b	3	GATCCATCCA TCCTTCCTC GATCCTGCCA CTGCAATC	X75550/1	67
wg1a9 (DXS1252)	2613 2614	4	CTTTTAGTGG GGTTACCTGG C CCTGAGTTAT CCCAGGAGCC	X73287	67
wg1c3	2740* 2789	5	CTAAAAATCT CACTTGATAC TTG AGCCCAACCG CTGCCTC	X76102	60
wg1c4 (D8S580)	2735 2790*	6	GAGTGAGACT CCATCTCAAC AGCTCTATCA GATTGGTTTG G*	X74778	60
wg1c5 (D14S299)	wg1c5a* wg1c5b	7	GATCTCAATA AACATTGATA CTGG CTGCATGAGC TAAAGCATAC TG	X73288	64
wg1d1 (D15S233)	2739 2811*	8	CACCACGTCC CGCCTCTAAT AGGTCTGGGT GACAGAACA	X75552	62

wg1d5 (D22S442)	2737 2791*	9	GACAGAGCGA CCTAGCTCCC	GACTCTGTCT A AGAACTTTCT CC	X74779	68
wg1d6 (D4S1639)	2649 2650	10	AGGAACTGTG TGCGGATTAG	GCTACAATAT TG CCTACTTCAT C	X73289	67
wg1d10 (D2S419)	wg1d10A wg1d10B	11	TAGCCTGCTT GGAGGGGGTC	ACCTTGAG ACTTTCTG	X74433	55
wg1d11	2746 2806	12	TTGGGCCTAG TGACAAAGCC	AAGGGAACTG AGACCTTG	X75553/4	62
wg1e1	2745 2794*	13	ATAAGAGAGT ACAGCAAGAC	AAATTTCTGG TG TCCACCAA	X75409	61
wg1e4 (D1S525)	wg1e4A* wg1e4B	14	TGACTGCAGC ATCATGGCAG	GGGATTAATC TACAGTCC	X74434	56
wg1e7 (D3S1749)	2744 2792	15	GTGACAGTGC CCCTCCGTCT	AAGTTTCTGC CTCCCTGC	X74780	60
wg1e12 (D7S822)	2743 2808	16	CTCCCTCCAC CATAGCCCAG	TCTGGAAATC TTTCTCAGGC	X75555	58
wg1f2 (D16S543)	wg1f2c* wg1f2d	17	GATCTTTAAC ACAGTGTGAG	TTAACTACAT TTGCAAG ACCCTGTCAA GG	X75556	62
wg1f4 (D6S483)	2702 2703	18	TATCCCTGAC CTGCCTCCTG	CCAGCAGCTG CCCTGAGGC	X73969	67

wg1g5 (D19S428)	wg1g5A wg1g5B*	19	GGGTGACAGA ATAAGACG ATGTCTATGC TGGAGACCTG	X74435	58
wg1g9 (D7S809)	wg1g9A wg1g9B*	20	GATCGTGCCA CTGTACC CGCGTGTCTA ACCTCATGGC	X73290	62
wg1h10 (D18S555)	2741* 2795	21	GTGCGATGGC AAAATAGATG ATTTTCTAGG AAAGAGCTAG C	X74781	60
wg2a5 (D4S1640)	wg2a5A wg2a5B*	22	CTCCAGCCTC TGTCTCC CTTGTTGGAA TGTGATGTGA AG	X73970	62
wg2b3 (D20S204)	wg2b3A wg2b3B*	23	CAGCCATCTG TTCAGTACTT TC TGATTTCCCA ACTGCAGACT TG	X74436	64
wg2c9 (D1S526)	wg2c9A* wg2c9B	24	TCAAGGCTGT GGTAAACTAT AATC ATAGACTCCT TTTTCTAAGC TCAG	X73291	64
wg0c5 (D3S1747)	2099 2100	25	AAAAAGAGCT AAAAGATGAA ACCG TGTTTCCAGT AACTTGAGTA ATTC	X73967	64
wg0f4 (D18S496)	2101 2102	26	ATGGTTTGTA AGGCAACAAA ACAC GCCCTTGGCA AGTGATGGAA G	X73214	64
wg0f5 (D2S413)	21032104	27	TCAGGAAACA TTGGGCTATT GTCC AGTGGCAAGA CTCATAGAAT TCCC	X73215	64
wg2e7	3297 3298	28	GCTTTGATGT ACTCATTGTC GATCTTTTTA AAAAAACTTT CC		60

wg2f7	3327 3328	29	ATGTTCTTTT ATGTTTGTTA ATATG TGACAGAGAC AGACCTTG		63
wg2f10	WG2F10A WG2F10B	30	CCATGAGGGT CTCTCTAGTC CAGTGAGCTG ACATCATG		60
wg2g4	WG2G4A WG2G4B	31	ACCTGAGGGA GCTCAAGAC GATCTGCCCG CCTCGG		66
wg2g12	WG2G12A WG2G12B	32	TCATGAAAGG ATATCCTGGG GAGCAGAACT CCATCCAC		62
wg2h11	WG2H11A WG2H11B	33	AGAGGGAGAC TTCATCAG GATCTGACCA CAAGGTAGTT G		62
wg3a6	WG3A6A WG3A6B	34	GTCATGATTT TGCCACTACT G ACATTTGCTA AGGGGATACA TC		64
wg3b10	3299 3300	35	TGAGTAACAA TCTCTGGCAG GATCCCTTTC ATTCCTTTCC		60
wg3f12	WG3F12A WG3F12B	36	GACAGATTGA GACCCCATC GCACAAAGCT GTGCATCAC		60
wg3h2	WG3H2A WG3H2B	37	TACCCAGTAG AAATACAACT GC CATGAAATCT AGTGCTAGAC AG		62

Table 4

Clone	Sequence Number(s)
wg0e7	1
wg1a2	2&3
wg1a3	4&5
wg1a9	6
wg1c3	7
wg1c4	8
wg1c5	9
wg1d1	10
wg1d5	11
wg1d6	12
wg1d10	13
wg1d11	14&15
wg1e1	16
wg1e4	17
wg1e7	18
wg1e12	19
wg1f2	20
wg1f4	21
wg1g5	22
wg1g9	23
wg1h10	24
wg2a5	25
wg2b3	26
wg2c9	27

wg0c5	28
wg0f4	29
wg0f5	30
wg2e7	31&32
wg2f7	33&34
wg2f10	35
wg2g4	36&37
wg2g12	38&39
wg2h11	40&41
wg3a6	42
wg3b10	43&44
wg3f12	45&46
wg3h2	47

CLAIMS

1. A method for the identification from DNA of a fragment comprising a simple tandem repeat locus comprising the steps of:

i) contacting a DNA library with at least one hybridisation probe so as to identify a population of DNA fragments enriched for simple tandem repeats;

ii) isolating and cloning said population; and

iii) screening of the resulting DNA library so as to identify an individual fragment comprising a simple tandem repeat locus.

2. A method according to claim 1 wherein the DNA library comprises a genomic DNA library.

3. A method according to either one of claims 1 and 2 wherein the DNA library comprises genomic human DNA fragments.

4. A method according to any one of claim 1 to 3 wherein the DNA library comprises subgenomic DNA fragments.

5. A method according to any one of the preceding claims wherein the average fragment size within the DNA library is less than about 1.5 kilobases.

6. A method according to any one of the preceding claims wherein the average fragment size within the DNA library is less than about 1 kilobase.

7. A method according to any one of the preceding claims wherein the average fragment size within the DNA library is from about 400bp to about 1000bp.

8. A method according to any one of the preceding claims wherein the hybridisation probe or probes is immobilised on a solid phase.

9. A method according to claim 7 wherein the solid phase comprises a nylon membrane.

10. A method according to any one of the preceding claims wherein the hybridisation probe or probes identifies a particular class of simple tandem repeats.

11. A method according to claim 9 wherein the class of simple tandem repeats is selected from the group of dimeric, trimeric, tetrameric, pentameric and hexameric tandem repeats.

12. A method according to any one of the preceding claims wherein the hybridisation probe or at least one of the hybridisation probes comprise a tandemly repeated region of greater than 200bp.

13. A method according to claim 12 wherein the probe or probes comprise repeats having at least 70% similarity to a given repeat sequence.

14. A method according to either one of claims 12 and 13 wherein the probe or probes comprise repeats having at least 80% similarity to a given repeat sequence.

15. A method according to any one of claims 12 to 14 wherein the probe or probes comprise repeats having at least 90% similarity to a given repeat sequence.

16. A method according to any one of the preceding claims wherein the hybridisation probes comprise a set of mixed trimeric or tetrameric repeat DNA.

17. A method according to any one of the preceding claims wherein the fragment or fragments comprising a simple tandem repeat locus so identified is subsequently amplified prior to cloning.

18. A method according to claim 17 wherein the amplification is effected by PCR.

19. A method according to claim 18 wherein universal linker sequences are ligated to the end or ends of the fragment or individual fragments.

20. A method according to either one of claims 18 and 19 wherein universal linker sequences are ligated to the end or ends of the fragment or individual fragments prior to the identification of the fragment or fragments.

21. A method according to either one of claims 19 and 20 wherein linker sequence specific primers are used to amplify the enriched population.

22. A method according to claim 21 wherein the linker sequences are removed subsequent to amplification and prior to cloning.

23. A method according to any one of the preceding claims for the identification from genomic DNA of a fragment comprising a simple tandem repeat locus comprising the steps of:

i) ligating universal linker sequences to the ends of fragments comprised in a genomic DNA library so as to form a library for PCR amplification;

ii) contacting said PCR library with at least one hybridisation probe so as to identify a population of library fragments enriched for simple tandem repeats;

iii) separating and amplifying said population by PCR; and

iv) cloning and screening the resulting amplification products so as to isolate an individual fragment comprising a simple tandem repeat locus.

24. A method according to any one of the preceding claims wherein screening is effected using at least one hybridisation probe comprising at least part of a simple tandem repeat.

25. A simple tandem repeat for use in a method of treatment or diagnosis of the human or animal body characterised in that it may be amplified at least in part by PCR using any one of pairs 1 to 37 of primers.

26. A simple tandem repeat according to claim 25 wherein it comprises at least the sequence of at least one of sequences 1 to 47.

27. A simple tandem repeat according to either one of claims 25 and 26 wherein it is polymorphic.

28. A simple tandem repeat according to any one of claims 25 to 27 wherein it has a heterozygosity of at least 80%.

29. A simple tandem repeat according to any one of claims 25 to 28 wherein it has a heterozygosity of at least 85%.

30. A simple tandem repeat according to any one of claims 25 to 29 wherein it has a heterozygosity of at least 90%.

31. A pair of amplification primers for use in a method of treatment or diagnosis of the human or animal body specific to any one of the simple tandem repeats of

any one of claims 25 to 30.

32. A pair of amplification primers according to claim 31 wherein they are PCR primers.

33. A probe for use in a method of treatment or diagnosis of the human or animal body specific to at least part of any one of the simple tandem repeats of any one of claims 25 to 30.

34. A method of genetic characterisation of the human or animal body wherein sample DNA is characterised by reference to at least one of the simple tandem repeats, primers and probes of any one of claims 25 to 33.

35. A method of genetic characterisation according to claim 34 wherein it comprises either the use of at least one pair of amplification primers or probe of any one of claims 31 to 33.

36. A method of genetic characterisation according to either one of claims 34 and 35 wherein it is a genetic mapping study.

Figure 1

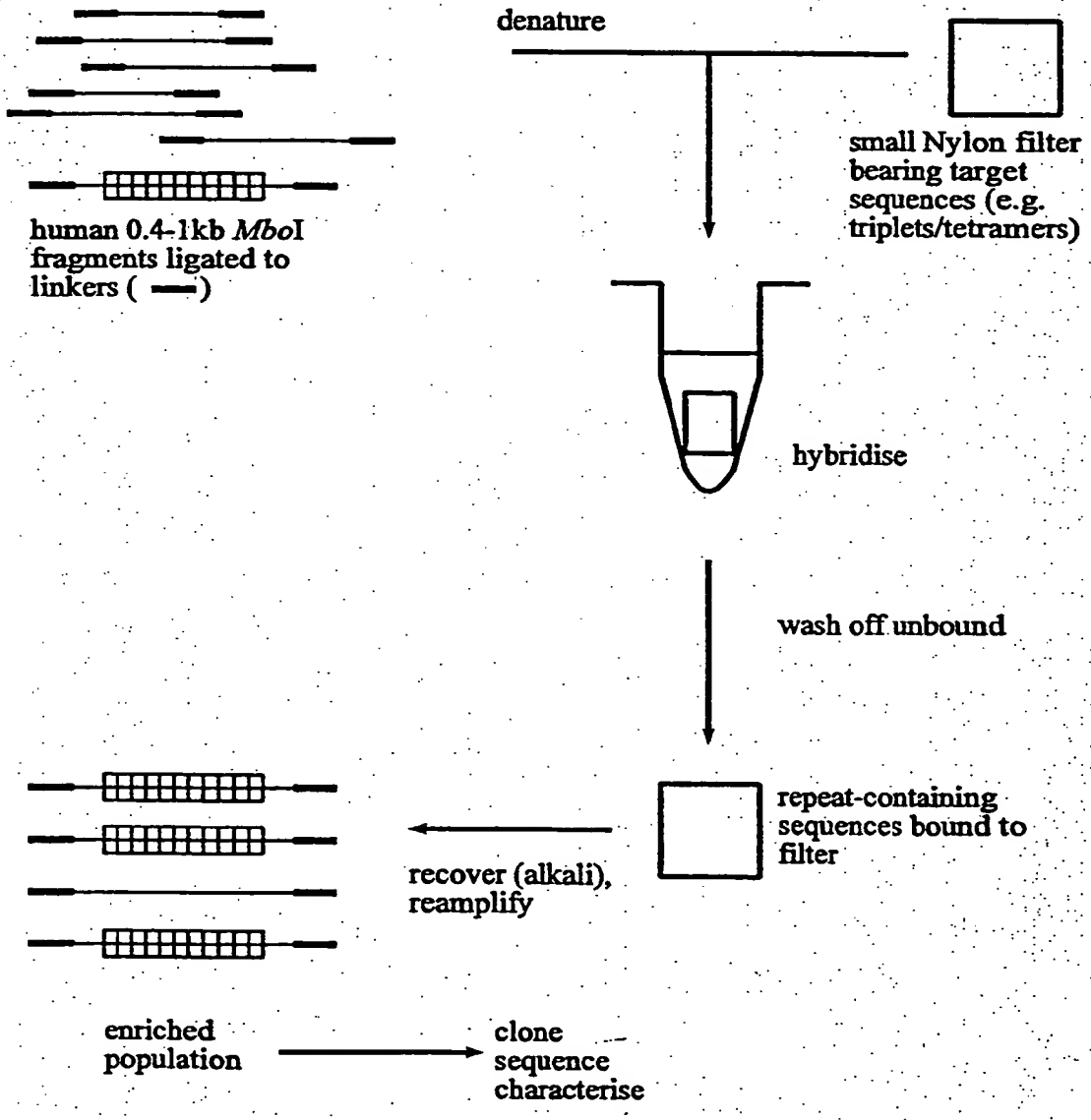


Figure 2

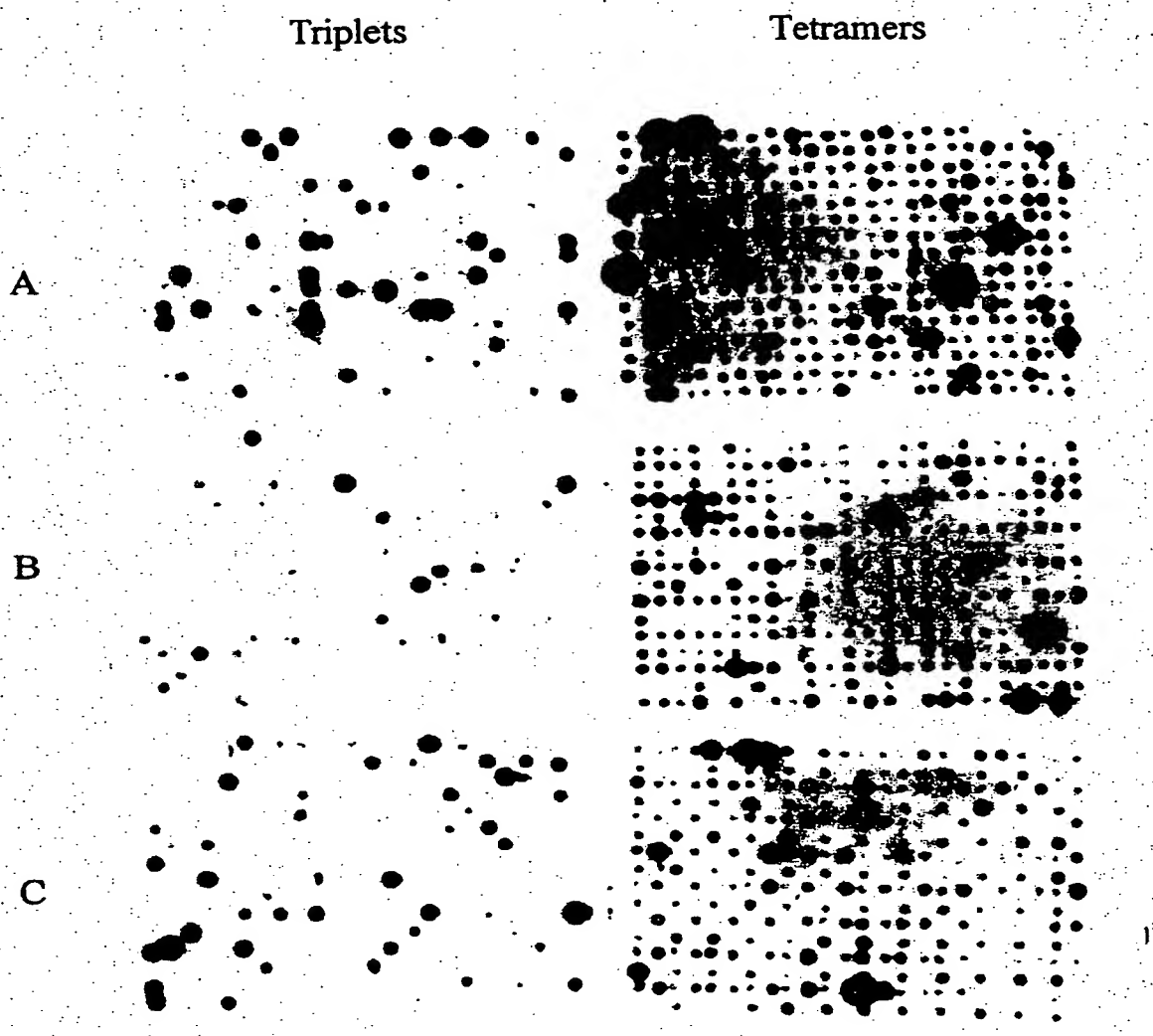


Figure 3

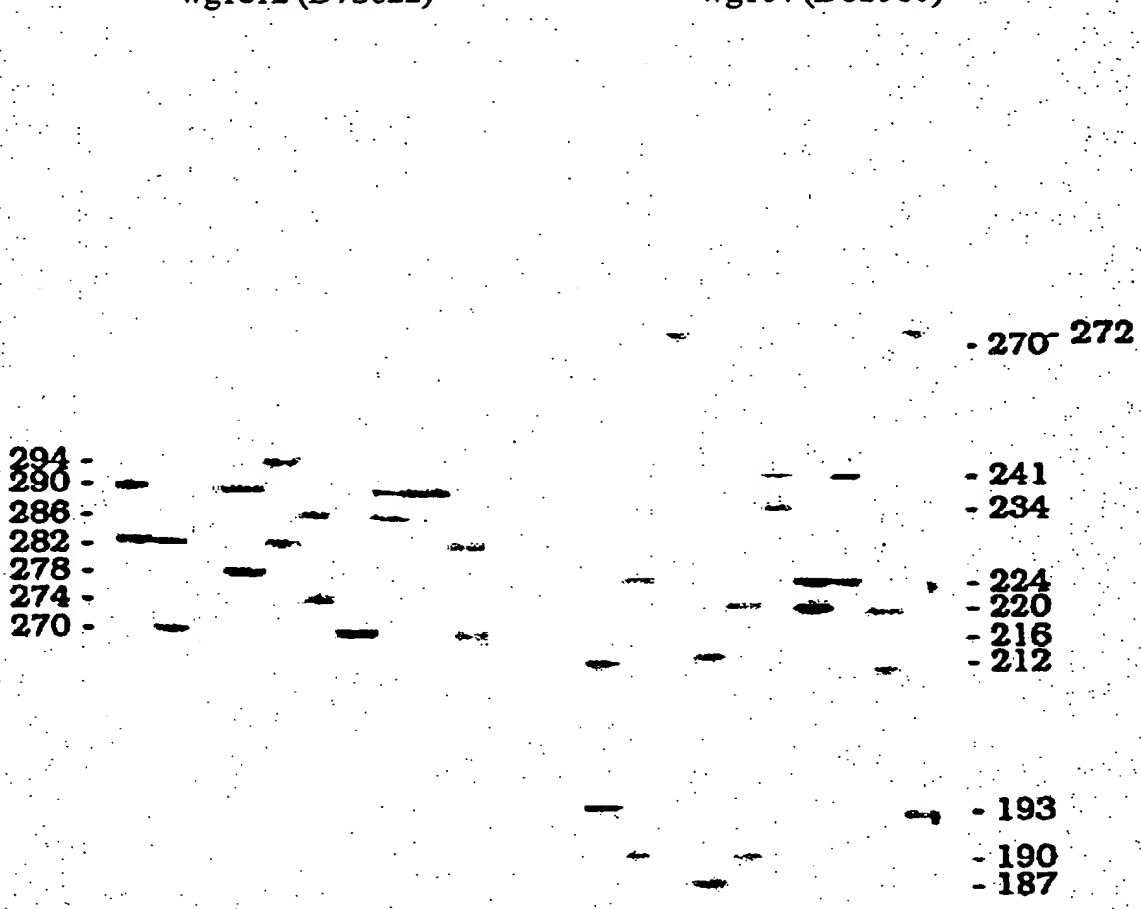

wg1e12 (D7S822)
wg1c4 (D8S580)
294
290
286
282
278
274
270
272
241
234
224
220
216
212
193
190
187

Figure 4